

Finite sample posterior concentration in high-dimensional regression

Nate Strawn, Artin Armagan, Rayan Saab*, Lawrence Carin, and David Dunson

*Department of Mathematics
Duke University*

e-mail: nstrawn@math.duke.edu; rayans@math.duke.edu

SAS Institute, Inc.
e-mail: artin.armagan@sas.com

*Department of Electrical and Computer Engineering
Duke University*
e-mail: lcarin@ee.duke.edu

*Department of Statistical Sciences
Duke University*
e-mail: dunson@stat.duke.edu

Abstract: We study the behavior of the posterior distribution in ultra high-dimensional Bayesian Gaussian linear regression models having $p \gg n$, with p the number of predictors and n the sample size. In particular, our focus is on obtaining non-asymptotic probabilistic bounds on the posterior probability assigned in neighborhoods of the true regression coefficient vector, β^0 , with these bounds used to study contraction of the posterior. We assume that β^0 is approximately S -sparse and obtain universal bounds via a Schwartz-type argument, though only well-structured priors exhibit good properties. Based upon these finite sample bounds, we examine the implied asymptotic contraction rates for several examples showing that sparsely-structured and heavy-tail shrinkage priors exhibit rapid contraction rates. Using brute force, we also demonstrate that a stronger result holds for the Uniform-Gaussian prior, which indicates that our main result can be strengthened and reinforces the fact that the estimates of the denominator in the Schwartz-type arguments are not sharp in the finite sample regime.

AMS 2000 subject classifications: Primary 62F15, 62F15; secondary 62F15.

Keywords and phrases: asymptotics, Bayesian, compressible prior, high-dimensional, posterior contraction, regression, shrinkage prior.

1. Introduction

Consider the Gaussian linear regression model

$$y_i = x_i^T \beta^0 + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1.1)$$

where x_i and β^0 are p -dimensional vectors. In modern applications, it has become common to collect ultra high-dimensional data in which the number of

*R. Saab was supported by a Banting Postdoctoral Fellowship, administered by the Natural Science and Engineering Research Council of Canada.

subjects n is much smaller than the number of predictors p . Assuming a prior distribution Π for the unknown coefficient vector β , our focus is on studying the concentration of the posterior distribution for β in neighborhoods of β^0 under the simplifying assumption that (1.1) is the true data-generating model. Although there is an increasingly rich literature studying the behavior of frequentist variable selection and point estimation in ultra high-dimensional settings [5, 13, 14], there has been essentially no work characterizing concentration of posterior distributions for β when $p > n$. By placing an appropriately-structured prior on β , one can potentially bypass the troubling issue of tuning parameter sensitivity while obtaining a useful probabilistic characterization of uncertainty. For example, the posterior can be used to construct credible regions for β .

In finite high dimensional settings, usual asymptotic justifications break down and it is important to carefully study frequentist properties of the posterior. There has been some relevant work in the literature. In [8], Ghosal obtained a Bernstein-von Mises theorem providing sufficient conditions to obtain asymptotic normality of the posterior distribution for β under model (1.1) allowing non-Gaussian residuals, but he requires p to grow much slower than n . Jiang [10] studied rates of convergence of the predictive distribution obtained using Bayesian variable selection within a generalized linear model having a diverging number of candidate predictors, but his results focus only on the predictive posterior of y given x and not on the posterior of β . More recently, Bontemps [2] obtained a Bernstein-von Mises theorem for a class of semiparametric and non-parametric Gaussian regression models. For the model (1.1) compared with [8], his results allow a faster growth rate of $p \leq n$. However, addressing our interest in $p \gg n$ requires new theory; in this much more challenging case, we do not attempt a Bernstein-von Mises result but instead provide explicit finite sample probabilistic bounds on the posterior probability assigned to neighborhoods of β_0 .

When $p \gg n$ there clearly needs to be some sort of dimensionality reduction or prior information included to make the problem tractable. One common flavor of dimensionality reduction corresponds to sparsity, which manifests in model (1.1) through assuming that only a small number of elements of the true coefficient vector β^0 are non-zero. Sparsity is a particularly natural assumption that appears implicitly in the literature of model selection through thresholding and shrinkage. Such model selection was originally performed in the $p < n$ case with the goal of surmounting overfitting and reducing the prediction error of a model [1]. The philosophy behind the more recent leap to $n \ll p$ motivated Donoho's rhetorical question: "Why go to so much effort to acquire all the data when most of what we get will be thrown away?" This philosophy gave birth to the field of compressed sensing [4, 7], which studies reconstruction and estimation of approximately sparse signals when $n \ll p$. Much of the rich literature on $n \ll p$ problems outside of the statistical community occurs within the compressed sensing community.

1.1. Contributions

In this paper, we provide theoretical validation of the Bayesian approach when $n \ll p$. Our main result, Theorem 3.1, employs a modification of Schwartz’s argument (the weapon of choice for Bayesian asymptotics) to exhibit an explicit bound on the expected concentration of a posterior for an arbitrary prior. The utility of this bound is evident when

- (i) the probability the prior assigns to a small ball around the true β_0 is not too small;
- (ii) the probability the prior assigns to signals that are not sparse (or approximately sparse) is very small.

These observations are embodied in Theorem 3.2, which summarizes conditions on a prior Π which ensure asymptotic posterior contraction and associated rates of contraction. To demonstrate the power of this result, we provide a number of worked examples including the Uniform-Gaussian, the Bernoulli-Gaussian, and Laplace priors. For these examples, bounds on the asymptotic rates of posterior contraction are derived.

We conclude by demonstrating that Theorem 3.1 is not sharp, and we discuss why this is the case. In particular, as we derived our main theorem using a Schwartz-type analysis, Theorem 4.1 indicates that the denominator estimate in the Schwartz-type analysis is too severe in the non-asymptotic setting.

1.2. Organization

In Section 2, we fix notation and provide background results that shall be employed throughout the paper. Section 3 introduces our main result, the explicit bound on expected posterior concentration for an arbitrary prior and a fixed problem size. We discuss the role of each term in this bound to indicate exactly when the bound is useful, and then provide a general theorem concerning asymptotic posterior contraction. Lastly, we apply our main result to calculate posterior contraction rates for some example priors. In Section 4, we discuss the sharpness of our main result and avenues for future exploration. In particular, we exhibit a theorem which provides a sharp guarantee for the Uniform-Gaussian prior. Appendices A and B contain the supporting technical material for Sections 3 and 4.

2. Preliminaries

2.1. Notation

Assuming model (1.1) is the true model with β^0 unknown, we focus on the posterior for observed data

$$y = X\beta^0 + e, \tag{2.1}$$

where y is the n -dimensional response, X is the $n \times p$ design matrix, $\Pi(\beta)$ is the prior on β^0 , and $e \sim \mathcal{N}(0, \sigma^2 I_n)$. For a fixed problem (S, n, p, X, β^0) , we designate the following assumptions:

- (A1) the i th column of X satisfies $\|X_i\|_{\ell_2}^2 = n$ for all $i = 1, \dots, p$
- (A2) β^0 is S -sparse ($\|\beta^0\|_{\ell_0} \leq S$)
- (A3) σ is known

Whenever we consider a sequence of problems, $(S_n, n, p_n, X(n), \beta^0(n))$ with $n \rightarrow \infty$, we additionally assume that $|\beta_i^0(n)| \leq C < \infty$ for all i and n and that the constants C and σ remain fixed as n increases. We shall let X^T denote the transpose of the matrix X , and we let $B_\varepsilon^{\ell_u}(\beta)$ denote the ℓ_u ball of radius ε centered at β .

The first essential condition for the success of $n \ll p$ analysis is that β^0 is S -sparse or S -compressible. To formalize these concepts, we recall the best k -term approximations and define the (S, R) -compressible vectors.

Definition 2.1. For any $\beta \in \mathbb{R}^p$ and any natural number $S \leq p$, let $\sigma_S(\beta)$ denote the best S -term approximation error of β so that

$$\sigma_S(\beta) = \inf_{\|\xi\|_{\ell_0} \leq S} \|\beta - \xi\|_{\ell_1}. \quad (2.2)$$

Furthermore, for any $R \geq 0$, let

$$\mathcal{P}_{S,R} = \{\beta \in \mathbb{R}^p : \sigma_S(\beta) \leq R\} \quad (2.3)$$

denote the set of (S, R) -compressible vectors.

Note that $\mathcal{P}_{S,0}$ is exactly the union of canonical S -dimensional subspaces in \mathbb{R}^p . When S and R are clear from the context, we shall simply let $\mathcal{P} = \mathcal{P}_{S,R}$.

Given the model (2.1), we let $f(y|\beta)$ denote the likelihood of $\beta \in \mathbb{R}^p$ given observed data $y \in \mathbb{R}^n$, and hence

$$f(y|\beta) = (2\pi\sigma^2)^{-n/2} \exp\{-\|y - X\beta\|_{\ell_2}^2/2\sigma^2\}. \quad (2.4)$$

For any $\beta \in \mathbb{R}^p$, Borel measurable $U \subset \mathbb{R}^n$, and Borel measurable function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, we let

$$\text{pr}_\beta(U) \text{ and } \mathbb{E}_\beta F \quad (2.5)$$

denote the probability of the event $y \in U$ given the parameter β and the expectation of $F(y)$ given the parameter β , respectively. We also let U^c denote the complement of the set U , and we let $\mathbf{1}_U$ denote the indicator function of U . Finally, for a linear operator $X : \mathbb{R}^p \rightarrow \mathbb{R}^n$, we use the notation $\|X\|_{\ell_u \rightarrow \ell_v}$ to denote the operator norm

$$\|X\|_{\ell_u \rightarrow \ell_v} = \max_{\{\beta : \|\beta\|_{\ell_u} = 1\}} \|X\beta\|_{\ell_v}. \quad (2.6)$$

2.2. The Dantzig selector

The Dantzig selector is an essential ingredient for our proofs. The properties of the Dantzig selector depend heavily upon the design matrix X , and one simple assumption laid out by Candès and Tao [5] is that the column norms of X all equal one. In our situation, it is convenient to rescale X so the column norms are all \sqrt{n} . We shall use $\tilde{X} = \frac{1}{\sqrt{n}}X$ as an intermediate quantity to translate the results of Candès and Tao to our setting.

Definition 2.2. For a response vector $y \in \mathbb{R}^n$ and a design matrix X , the Dantzig selector is the solution to the program

$$\min \|\beta\|_{\ell_1} \text{ subject to } \|\tilde{X}^T(y - \tilde{X}\beta)\|_{\ell_\infty} \leq \lambda_p \sigma \quad (2.7)$$

where $\lambda_p = \sqrt{2(1+\alpha)\log p}$. The role of the free parameter $\alpha > 0$ is made apparent in Theorem 2.1. We let $\tilde{\beta}$ denote the solution to this linear programming problem, and set $\hat{\beta} = \tilde{\beta}/\sqrt{n}$.

In order to ensure reconstruction properties for the Dantzig selector for all β of a sufficient sparsity, we must put conditions on \tilde{X} . The first quantity of interest is the restricted isometry constant, which is the smallest constant $\delta_k(\tilde{X})$ satisfying

$$(1 - \delta_k)\|b\|_{\ell_2}^2 \leq \|\tilde{X}b\|_{\ell_2}^2 \leq (1 + \delta_k)\|b\|_{\ell_2}^2 \quad (2.8)$$

for all $b \in \mathcal{P}_{k,0}$. Ideally, the constant δ_k is small enough to ensure that sufficiently sparse b are far from the kernel of \tilde{X} . The other quantity of interest is the restricted orthogonality constant $\theta_{k,k'}(\tilde{X})$, which is defined to be the smallest constant such that

$$|\langle \tilde{X}_T b, \tilde{X}_{T'} b' \rangle| \leq \theta_{k,k'} \|b\|_{\ell_2} \|b'\|_{\ell_2} \quad (2.9)$$

for all b, b' , disjoint $T, T' \subset \{1, 2, \dots, p\}$, where $\tilde{X}_T, \tilde{X}_{T'}$ are the matrices formed by concatenating the columns of \tilde{X} with indices in T and T' respectively, $|T| \leq k$, $|T'| \leq k'$, and $|T| + |T'| \leq p$. Again, the ideal $\theta_{k,k'}$ is small, so disjoint collections of columns of \tilde{X} span nearly orthogonal subspaces.

With the restricted isometry and restricted orthogonality constants defined, we are now able to translate the theorem of Candès and Tao into our setting.

Theorem 2.1 (Candès and Tao '05). *Let S be fixed so that $\delta_{2S}(\tilde{X}) + \theta_{S,2S}(\tilde{X}) < 1$ and fix $R \geq 0$. If $\beta^0 \in \mathcal{P}_{S,R}$, then the rescaled solution to (2.7) satisfies*

$$\|\hat{\beta} - \beta^0\|_{\ell_2} \leq \frac{4\sqrt{2}\sigma}{1 - \delta - \theta} \sqrt{\frac{2(1+\alpha)S \log p}{n}} + 2 \frac{1 - \delta + \theta}{1 - \delta - \theta} \frac{R}{\sqrt{S}} \quad (2.10)$$

with probability greater than $1 - \frac{1}{p^\alpha \sqrt{\pi \log p}}$.

For completeness, we prove this version of the theorem in Appendix B. Since we shall employ the above condition on \tilde{X} to invoke this theorem and to perform further analysis, we add the following assumption to our arsenal.

(A4) $\delta \equiv \delta_{2S}(\tilde{X})$ and $\theta \equiv \theta_{S,2S}(\tilde{X})$ satisfy $\delta + \theta < 1$

In the case of increasing problem sizes, we shall assume that δ and θ remain fixed (or are at least nonincreasing as the problem size increases). While at first glance this may seem to constrain the applicability of our theory, such conditions are standard in the theoretical literature on sparse reconstructions and obtaining universal statements in problems where $n \ll p$ without similar conditions is an open problem. Another possible concern is that verification of these constants is combinatorially complex, however it has been well established that many families of random matrices satisfy this condition with high probability when $n \geq CS(\log p)^c$, for some $c \geq 1$. In particular, matrices whose entries are drawn i.i.d. $\mathcal{N}(0, 1/n)$, and matrices with sub-Gaussian entries satisfy this condition with high probability with $c = 1$. Other matrices that satisfy such a condition (albeit with $c > 1$) include $n \times p$ matrices whose rows are drawn (uniformly) at random from orthonormal bases such as the Fourier basis. The interested reader is referred, for example, to [4].

3. Main result and examples

Theorem 3.1. *Suppose $\beta^0 \in \mathcal{P} = \mathcal{P}_{S,R}$ and that Π is an arbitrary prior on \mathbb{R}^p . Let $\mathcal{B} = \{\beta \in \mathbb{R}^p : \|\beta - \beta^0\|_{\ell_2} > 2\varepsilon\}$, with*

$$\varepsilon = \frac{8\sigma}{1 - \delta - \theta} \sqrt{\frac{(1 + \alpha)S \log p}{n}} + 2 \frac{1 - \delta + \theta}{1 - \delta - \theta} \frac{R}{\sqrt{S}}, \quad (3.1)$$

and assume (A1), (A3) and (A4). For any $\alpha > 0$, $\kappa > 0$, $0 < \nu < \alpha$, and all u and v satisfying $1/u + 1/v = 1$ with $u \geq 1$,

$$\mathbb{E}_{\beta^0} \Pi(\mathcal{B}|y) \leq \frac{1}{p^\alpha \sqrt{\pi \log p}} \quad (3.2)$$

$$+ \frac{\Pi(\mathcal{B} \setminus \mathcal{P})}{\Pi(\mathcal{D}_{\nu,\kappa}) p^{-\nu}} \quad (3.3)$$

$$+ \frac{1}{\Pi(\mathcal{D}_{\kappa,\nu}) p^{\alpha-\nu} \sqrt{\pi \log p}} \quad (3.4)$$

$$+ p r_{\beta^0}(\mathcal{A}_\kappa^c), \quad (3.5)$$

where

$$\mathcal{D}_{\nu,\kappa} = \{\beta \in \mathbb{R}^p : \|\beta - \beta^0\|_{\ell_u} < C_{\nu,\kappa}\}, \quad (3.6)$$

$$C_{\nu,\kappa} = \left(\frac{\sqrt{2\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2 \nu \log p}}{\kappa + \sqrt{\kappa^2 + 2\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2 \nu \log p}} \right) \sqrt{\frac{2\sigma^2 \nu \log p}{\|X^T X\|_{\ell_u \rightarrow \ell_v}}}, \quad (3.7)$$

and

$$\mathcal{A}_\kappa = \{y \in \mathbb{R}^n : \|X^T(y - X\beta^0)\|_{\ell_v} > \kappa\}. \quad (3.8)$$

Ultimately, we want to apply this theorem to obtain concentration bounds for large classes of sparsity promoting priors. To do this we must understand the contribution of each individual term of the inequality stated in Theorem 3.1. Immediately, we may observe that the terms (3.2) and (3.5) are independent of Π . The term in (3.2) comes from using the Dantzig estimator in our hypothesis test. As such, we have little control over this term aside from adjusting the α parameter. The fourth term, (3.5) is controlled by the parameter κ , and measures the tail probability of a random vector's magnitude. With the simple choice of $u = 1$ and $v = \infty$, we may set $\kappa = \sqrt{n}\lambda_p\sigma$ and bound this term by $(p^\alpha\sqrt{\pi\log p})^{-1}$. For all of our examples, we adopt this strategy to essentially eliminate (3.5) and add a coefficient of 2 to the term (3.2).

Having discussed the terms that are independent of the prior, we turn our attention to the middle terms. The term (3.4) depends inversely upon $\Pi(\mathcal{D}_{\nu,\kappa})$, the probability the prior assigns to a ball with a very small radius around β^0 . The behavior of this term illustrates the role that sparsity plays in the behavior of the posterior. To see this note, that in order to control this term, we must increase α . However, if $\Pi(\mathcal{D}_{\nu,\kappa})$ is proportional to the volume of $\mathcal{D}_{\nu,\kappa} \subset \mathbb{R}^p$, then α must overcome p , and

$$\varepsilon \propto \sqrt{\frac{S(1 + Cp)\log p}{n}},$$

may be quite large. On the other hand, a sparsity promoting prior can rescue the asymptotic behavior of the posterior. Because a sparsity promoting prior is concentrated very near S -dimensional subspaces, the probability assigned to a small ball around a sparse vector is proportional to the volume of a ball in \mathbb{R}^S . Thus, α can remain $O(S)$, and ε shrinks asymptotically if $S^2 \log p$ is $o(n)$.

Finally, we discuss the term (3.3). The presence of $\Pi(\mathcal{B} \setminus \mathcal{P})$ in the numerator indicates that this term can only be controlled if the prior encourages sparsity. The presence of the term $\Pi(\mathcal{D}_{\nu,\kappa})$ in the denominator then indicates that this term can only be controlled if the prior encourages sparse β . In particular, if Π is a compressible prior (see [6, 9]), $\Pi(\mathcal{B} \setminus \mathcal{P})$ should be small. In general, the decay of $\Pi(\mathcal{B} \setminus \mathcal{P})$ must overcome the growth of a p^S term produced by $\Pi(\mathcal{D}_{\nu,\kappa})$.

Based on Theorem 3.1, we may exhibit a general posterior contraction result depending upon $\Pi(\mathcal{B} \setminus \mathcal{P})$ and $\Pi(\mathcal{D}_{\nu,\kappa})$.

Theorem 3.2. *Suppose $(S_n, n, p_n, X(n), \beta^0(n))$ is a sequence of problems satisfying (A1) through (A4), and that Π_n is a sequence of priors on \mathbb{R}^{p_n} such that*

- i. $\Pi(\mathcal{D}_n) \geq p_n^{-\eta_n}$
- ii. $\Pi(\mathcal{B}_n \setminus \mathcal{P}) \leq p_n^{-\phi_n}$

where

$$\mathcal{D}_n = \left\{ \beta \in \mathbb{R}^{p_n} : \|\beta - \beta^0(n)\|_{\ell_1} < \sqrt{\frac{\sigma^2 \log p_n}{2(2 + \alpha)n}} \right\}. \quad (3.9)$$

Let $\mathcal{B}_n = \{\beta \in \mathbb{R}^p : \|\beta - \beta^0(n)\|_{\ell_2} > 2\varepsilon_n\}$ where

$$\varepsilon_n = \frac{8\sigma}{1 - \delta - \theta} \sqrt{\frac{(1 + \alpha_n)S_n \log p_n}{n}}. \quad (3.10)$$

Then

$$\mathbb{E}_{\beta^0} \Pi_n(\mathcal{B}_n | y) \leq (2 + p_n^{\eta_n+1}) \frac{1}{p_n^{\alpha_n} \sqrt{\pi \log p_n}} + p_n^{-\phi_n + \eta_n + 1}. \quad (3.11)$$

Example 3.1. First, we turn our attention to a case that admits the simplest (but still somewhat involved) analysis. We let $\{0, 1\}_S^p$ denote the p -length binary sequences with exactly S nonzero entries and fix the model

$$\beta_i \sim \gamma_i \mathcal{N}(0, V^2) + (1 - \gamma_i) \delta_0 \quad (3.12)$$

$$\gamma \sim \text{Uniform}(\{0, 1\}_S^p) \quad (3.13)$$

where $\text{Uniform}(\{0, 1\}_S^p)$ is the distribution with equal $(1/\binom{p}{S})$ probability for all $\gamma \in \{0, 1\}_S^p$. First, note that $\Pi(\mathcal{B} \setminus \mathcal{P}_S) = 0$, which eliminates the term (3.3). Next, set $u = 1$, $v = \infty$, and $\kappa = \sqrt{2(1 + \alpha)n\sigma^2 \log p}$. This ensures that $\text{pr}_{\beta^0}(\mathcal{A}_\kappa^c) \leq \frac{1}{p^\alpha \sqrt{\pi \log p}}$. For simplicity, we assume that $\nu = 1$ and obtain following reduction from Theorem 3.1:

$$\mathbb{E}_{\beta^0} \Pi(\mathcal{B} | y) \leq \left(2 + \frac{p}{\Pi(\mathcal{D}_{\nu, \kappa})}\right) \frac{1}{p^\alpha \sqrt{\pi \log p}} \quad (3.14)$$

We now only need to estimate $\Pi(\mathcal{D}_{\nu, \kappa})$. To that end, suppose that β^0 has support T and denote by $\text{Vol}(\mathcal{D}_{\nu, \kappa}^S)$ the volume of an S -dimensional ℓ_1 -ball with the same radius as $\mathcal{D}_{\nu, \kappa}$. Let M denote the minimum of $\prod_{i \in T} \mathcal{N}(\beta_i | 0, V^2)$ over $\mathcal{D}_{\nu, \kappa}^S$. Then

$$\Pi(\mathcal{D}_{\nu, \kappa}) \geq M \text{Vol}(\mathcal{D}_{\nu, \kappa}^S) \Pi(\gamma = \mathbf{1}_T) = \left(\frac{p}{S}\right)^{-1} M \text{Vol}(\mathcal{D}_{\nu, \kappa}^S) \quad (3.15)$$

$$\geq \left(\frac{p}{S}\right)^{-1} (2\pi V^2)^{-S/2} e^{-\|\beta^0\|_{\ell_2}^2 / 2V^2} e^{-C_{\nu, \kappa}^2 / 2V^2} \frac{(2C_{\nu, \kappa})^S}{\Gamma(1 + S)} \quad (3.16)$$

$$\geq \frac{e^{-\|\beta^0\|_{\ell_2}^2 / 2V^2 - C_{\nu, \kappa}^2 / 2V^2}}{\sqrt{2\pi(eV)^2}^S} \left(\frac{2C_{\nu, \kappa}}{p}\right)^S \quad (3.17)$$

Since $\|X^T X\|_{\ell_1 \rightarrow \ell_\infty} = \max_{ij} |(X^T X)_{ij}| = n$ (see, e.g., [12]), we have

$$\begin{aligned} C_{\nu, \kappa} &= \left(\frac{\sqrt{2n\sigma^2 \log p}}{\sqrt{2n\sigma^2(1 + \alpha) \log p} + \sqrt{2n\sigma^2(1 + \alpha) \log p + 2n\sigma^2 \log p}} \right) \sqrt{\frac{2\sigma^2 \log p}{n}} \\ &= \sqrt{\frac{\sigma^2 \log p}{2(2 + \alpha)n}} \end{aligned} \quad (3.18)$$

Combining (3.17) and (3.18), we have

$$\begin{aligned}\Pi(\mathcal{D}_{\nu,\kappa}) &\geq e^{-C_{\nu,\kappa}^2/2V^2} \left(\frac{\sqrt{2\sigma^2}}{e^{C^2/2V^2} \sqrt{\pi(eV)^2}} \right)^S \left(\frac{1}{p} \sqrt{\frac{\log p}{(2+\alpha)n}} \right)^S \\ &= \eta_0(\eta_1)^S \left(\frac{1}{p} \sqrt{\frac{\log p}{(2+\alpha)n}} \right)^S\end{aligned}\quad (3.19)$$

where we have set $\eta_0 = e^{-C_{\nu,\kappa}^2/2V^2}$ and $\eta_1 = e^{-C^2/2V^2} \sqrt{\frac{2\sigma^2}{\pi e^2 V^2}}$. Note that η_1 is constant, and (though it depends on n, p , and α) η_0 is approximately constant in the asymptotic regime. Combining (3.19) with (3.14), we have the bound

$$\mathbb{E}_{\beta^0} \Pi(\mathcal{B}|y) \leq \left(2 + \frac{p}{\eta_0} \left(\frac{p}{\eta_1} \sqrt{\frac{(2+\alpha)n}{\log p}} \right)^S \right) \frac{1}{p^\alpha \sqrt{\pi \log p}}. \quad (3.20)$$

Now, consider a sequence of problems $(S_n, n, p_n, X(n), \beta^0(n))$ satisfying (A1) through (A4), and suppose we employ the Uniform-Gaussian prior with parameter V fixed for each n . Then the bound in (3.20) applies to $\mathbb{E}_{\beta^0(n)} \Pi(\mathcal{B}_n|y(n))$ for each n , where the radius of \mathcal{B}_n is $2\varepsilon_n$ with

$$\varepsilon_n = \frac{8\sigma}{1-\delta-\theta} \sqrt{\frac{(1+\alpha_n)S_n \log p_n}{n}}. \quad (3.21)$$

The most problematic contribution to the bound in (3.20) is $p_n^{S_n}$, but we may adjust α_n so that $p_n^{\alpha_n}$ overcomes this term asymptotically. Thus, in order to obtain asymptotic consistency, we require $\alpha_n - S_n \rightarrow \infty$ and $(1+\alpha_n)S_n \log p_n = o(n)$. This is possible if we set $\alpha_n = S_n \log p_n$ and assume $S_n \log p_n = o(\sqrt{n})$.

Example 3.2. Now, we assume the model

$$\beta_i \sim \gamma_i \mathcal{N}(0, V^2) + (1 - \gamma_i) \delta_0 \quad (3.22)$$

$$\gamma_i \sim \text{Bernoulli}(\phi) \quad (3.23)$$

where $\phi \in (0, 1)$ controls the sparsity of the prior. As in the Uniform-Gaussian prior, we shall set $u = 1$, $v = \infty$, $\nu = 1$, and $\kappa = \sqrt{2(1+\alpha)n\sigma^2 \log p}$. In order to carry out a meaningful analysis in this case, we must assume that β^0 is K -sparse and that $p\phi = K$. By Chernoff-Hoeffding, we have that

$$\text{pr} \left\{ \sum \gamma_i \geq S \right\} \leq \left(\frac{K}{S} \right)^S \left(\frac{p-K}{p-S} \right)^{p-S}. \quad (3.24)$$

Note that this is a bound for $\Pi(\mathcal{B} \setminus \mathcal{P}_{S,\cdot})$. We are left with producing an estimate

for $\Pi(\mathcal{D}_{\nu,\kappa})$:

$$\Pi(\mathcal{D}_{\nu,\kappa}) = \sum_{\gamma} \Pi(\mathcal{D}_{\nu,\kappa}|\gamma) \Pi(\gamma) \quad (3.25)$$

$$= \sum_{k=0}^{p-K} \binom{p-K}{k} \phi^{K+k} (1-\phi)^{p-K-k} \Pi(\mathcal{D}_{\nu,\kappa}|\gamma) \quad (3.26)$$

$$\begin{aligned} &\geq \phi^K \sum_{k=0}^{p-K} \binom{p-K}{k} \phi^k (1-\phi)^{p-K-k} \frac{e^{-\frac{\|\beta^0\|_2^2}{2V^2} - \frac{C_{\nu,\kappa}^2}{2V^2}} (2C_{\nu,\kappa})^{K+k}}{\sqrt{2\pi V^2}^{K+k} \Gamma(1+K+k)} \\ &\geq \eta_0 \left(\frac{2\phi C_{\nu,\kappa}}{e^{C^2/2V^2} \sqrt{2\pi V^2}} \right)^K \sum_{k=0}^{p-K} \binom{p-K}{k} \left(\frac{2\phi C_{\nu,\kappa}}{\sqrt{2\pi V^2}} \right)^k (1-\phi)^{p-K-k} \frac{1}{(K+k)!} \\ &\geq \eta_0 \left(\frac{2\phi C_{\nu,\kappa}}{K \sqrt{2\pi V^2}} \right)^K \sum_{k=0}^{p-K} \binom{p-K}{k} \left(\frac{2\phi C_{\nu,\kappa}}{p \sqrt{2\pi V^2}} \right)^k (1-\phi)^{p-K-k} \\ &= \eta_0 \left(\frac{\eta_1}{p} \frac{S}{K} \sqrt{\frac{\log p}{(2+\alpha)n}} \right)^K \left(1 - \frac{K}{p} + \frac{\eta_1}{p} \frac{K}{p} \sqrt{\frac{\log p}{(2+\alpha)n}} \right)^{p-K}. \end{aligned} \quad (3.27)$$

Here, η_0 is as in the previous example and $\eta_1 = e^{-C/2V^2} \sqrt{\frac{2\sigma^2}{\pi V^2}}$. In this case, the term (3.3) in Theorem 3.1 is bounded by

$$\frac{p}{\eta_0} \left(\frac{K}{S} \right)^{S+K} \left(\frac{p}{\eta_1} \sqrt{\frac{\log p}{(2+\alpha)n}} \right)^K \left(\frac{p-K}{p-S} \right)^{p-S} \left(1 - \frac{K}{p} + \frac{\eta_1}{p} \frac{K}{p} \sqrt{\frac{\log p}{(2+\alpha)n}} \right)^{K-p}, \quad (3.28)$$

Now, consider a sequence of problems $(K_n, n, p_n, X(n), \beta^0(n))$ satisfying (A1) through (A4), and suppose we employ the Bernoulli-Gaussian prior with parameters V and $\phi_n = K_n/p_n$ for each n . In order to handle the term (3.4), we need to choose α_n so that $\alpha_n - K_n \rightarrow \infty$. Thus, we set $\alpha_n = K_n \log p_n$. In order to deal with the term (3.3), we require $S_n - K_n \log p_n \rightarrow \infty$. Finally, to shrink the radius of \mathcal{B}_n , which is twice

$$\varepsilon_n = \frac{8\sigma}{1-\delta-\theta} \sqrt{\frac{(1+\alpha_n)S_n \log p_n}{n}}, \quad (3.29)$$

we may assume that $\alpha_n = K_n \log p_n$, $S_n = K_n \log^2 p_n$, and thus we need $K_n \log^2 p_n = o(\sqrt{n})$.

Example 3.3. Now, we examine the concentration of the posterior under a Laplace prior. First, we set

$$\phi = \frac{\lambda}{2} \int_{-R}^R e^{-\lambda|x|} dx \quad (3.30)$$

and obtain the bound

$$\Pi(\mathcal{B} \setminus \mathcal{P}_{S,R}) \leq \Pi(\mathcal{P}_{S,R}^c) \quad (3.31)$$

$$\leq \left(\frac{p\phi}{S}\right)^S \left(\frac{p(1-\phi)}{p-S}\right)^{p-S}. \quad (3.32)$$

On the other hand, we have

$$\Pi(\mathcal{D}) \geq e^{-\lambda\|\beta^0\|_{\ell_1}} \Pi\left(B_{C_{\nu,\kappa}}^{\ell_2}(0)\right) \quad (3.33)$$

$$= e^{-\lambda\|\beta^0\|_{\ell_1}} \frac{1}{\Gamma(p+1)} \left(\frac{\lambda}{2}\right)^p \int_0^{C_{\nu,\kappa}} (2r)^p e^{-\lambda r} dr \quad (3.34)$$

$$= e^{-\lambda\|\beta^0\|_{\ell_1}} \frac{\gamma(p+1, \lambda C_{\nu,\kappa})}{\lambda \Gamma(p+1)}. \quad (3.35)$$

From this, we arrive at the bound

$$\mathbb{E}_{\beta^0} \Pi(\mathcal{B}|y) \leq \left(1 + e^{\lambda\|\beta^0\|_{\ell_1}} \frac{\lambda p \Gamma(p+1)}{\gamma(p+1, \lambda C_{\nu,\kappa})}\right) \frac{2}{p^\alpha \sqrt{\pi \log p}} \quad (3.36)$$

$$+ \frac{\lambda p e^{\lambda\|\beta^0\|_{\ell_1}} \Gamma(p+1)}{\gamma(p+1, \lambda C_{\nu,\kappa})} \left(\frac{p\phi}{S}\right)^S \left(\frac{p(1-\phi)}{p-S}\right)^{p-S}. \quad (3.37)$$

In both of these terms, we must overcome the $\Gamma(p+1) = p!$ factor. While the second term can be handled by scaling λ appropriately, getting the first term to decay requires $\alpha_n \sim O(p_n)$, which means that n must overcome $p_n \log p_n$. This is clearly impossible if we are interested in the $n \ll p$ regime. Since these bounds are sharp, we must conclude that Theorem 3.1 is not powerful enough to reveal any contraction behavior for the posterior under a Laplace prior. This of course does not mean that the Laplace prior does not concentrate, but that our theory cannot say anything about its concentration.

4. Evidence supporting a sharper result

Theorem 3.1 is the first result of its kind, but we are able to construct examples which behave provably better than what the Schwartz-type analysis can yield. In particular, with a little effort we may demonstrate the following concentration result for the Uniform-Gaussian prior.

Theorem 4.1. *Assume (A1) through (A4) and that Π is the Uniform-Gaussian prior with parameters S and V . Fix $\alpha > \max\left\{0, \frac{-29-32\theta+31\delta}{30+32\theta-32\delta}\right\}$ and let*

$$\varepsilon = \frac{C_1 \sigma^2 \sqrt{S}/n + (C_2 \sigma V^2 + C_3 \sigma^2/n) \sqrt{\frac{(1+\alpha)S \log p}{n}}}{(1-\delta)V^2 + \sigma^2/n}, \quad (4.1)$$

where the positive constants C_1 , C_2 and C_3 depend only δ and θ . If $(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2})\varepsilon^2 \geq S/2$, then there exists a constant $\eta = \eta(\alpha, \delta, \theta) > 0$ so that

$$\Pi(B_{2\varepsilon}^{\ell_2}(\beta^0)|y) \geq \frac{1 - e^{-\frac{1}{4}(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2})\varepsilon^2}}{1 + \left(e^{2\frac{n(1+\delta)V^2 + \sigma^2}{n(1-\delta)V^2 + \sigma^2}}\right)^{S/2} e^{\frac{1}{n(1-\delta)V^2 + \sigma^2} \frac{\|y\|_{\ell_2}^2}{2}} S^{-S} p^{-\eta S}} \quad (4.2)$$

with probability greater than $1 - 1/p^\alpha \sqrt{\pi \log p}$ on the draw of y .

Comparing this bound with that given in Example 3.1, it is clear that this theorem is much sharper. In particular, note that we no longer need to scale α to obtain asymptotic contraction. A very crude approximation in the asymptotic regime would be $\varepsilon \approx \sqrt{\frac{S \log p}{n}}$ and

$$\Pi(B_{2\varepsilon}^{\ell_2}(\beta^0)|y) \approx \frac{1 - Q_1 p^{-\eta_1 S}}{1 + Q_2 p^{-\eta_2 S}} \quad (4.3)$$

with probability exceeding $1 - 1/p^\alpha \sqrt{\pi \log p}$. In order to obtain contraction, we simply let $S \log p = o(n)$. This is as good a result as one can hope for, as (depending on X) n must be at least $CS \log(p/S)$ to guarantee (A4).

Why doesn't Theorem 3.1 provide us as sharp a guarantee? While Schwartz-type arguments are perfect for asymptotic situations where the likelihood is concentrating very heavily on a small ball, in a finite undersampling regime the reduction of the denominator term to a small ball is too dramatic, and excludes a large portion of the integral. In particular, focusing on a small ball of the Uniform-Gaussian prior introduces the term $\binom{p}{S}$ and when we approach the term

$$\frac{\Pi(\mathcal{B} \setminus \mathcal{P})}{\Pi(\mathcal{D}_{\nu, \kappa})}, \quad (4.4)$$

the shrinking radius of the ball $\mathcal{D}_{\nu, \kappa}$ forces us to require a suboptimal sparsity level.

These observations and Theorem 4.1 suggest that a stronger result should hold. In particular, the only pertinent restriction the prior should obey is that $\Pi(\mathcal{P}^C)$ should be sufficiently small. This condition forces the posterior to concentrate on the set of sparse β and the likelihood automatically concentrates on the sparse solutions to the linear inverse problem.

Appendix A

The proof of our main result is a modification of the argument originally devised by Schwartz [11]. In order to employ her strategy, we first find a large set of y 's for which the numerator of $\Pi(\beta|y)$ admits a controllable upper bound, and then we find another large set of y 's for which the denominator admits a controllable lower bound.

In the literature, this former set is identified with a hypothesis test which enjoys strong consistency behavior. As is often the case, we may base this hypothesis test on a frequentist estimator, and our estimator of choice is the Dantzig selector and we employ Theorem 2.1 to exploit the theoretical properties of the Dantzig selector. The theoretical properties of the LASSO estimator [3?] could also be exploited to form such a hypothesis test.

Proof of Theorem 2.1. Let $\beta = \sqrt{n}\beta^0$, set $h = \tilde{\beta} - \beta$, and suppose T_0 and T_{01} follow the precedent set in [5]. First, we note that

$$\|h_{T_0^c}\|_{\ell_1} \leq \|h_{T_0}\|_{\ell_1} + 2\|\beta_{T_0^c}\|_{\ell_1}. \quad (4.5)$$

By Lemma 3.1 of [5], we then have

$$\|h\|_{\ell_2} \leq \|h_{T_{01}}\|_{\ell_2} + S^{-1/2}\|h_{T_0^c}\|_{\ell_1} \quad (4.6)$$

$$\leq \|h_{T_{01}}\|_{\ell_2} + S^{-1/2}(\|h_{T_0}\|_{\ell_1} + 2\|\beta_{T_0^c}\|_{\ell_1}) \quad (4.7)$$

$$\leq \|h_{T_{01}}\|_{\ell_2} + \|h_{T_0}\|_{\ell_2} + 2S^{-1/2}\|\beta_{T_0^c}\|_{\ell_1} \quad (4.8)$$

$$\leq 2\|h_{T_{01}}\|_{\ell_2} + 2S^{-1/2}\|\beta_{T_0^c}\|_{\ell_1}. \quad (4.9)$$

Moreover, Lemma 3.1 also gives us

$$\|h_{T_{01}}\|_{\ell_2} \leq \frac{1}{1-\delta}\|\tilde{X}_{T_{01}}^T \tilde{X} h\|_{\ell_2} + \frac{\theta}{1-\delta}S^{-1/2}\|\beta_{T_0^c}\|_{\ell_1} \quad (4.10)$$

$$\leq \frac{2\sqrt{2}}{1-\delta}S^{1/2}\lambda_p + \frac{\theta}{1-\delta}S^{-1/2}(\|h_{T_0}\|_{\ell_1} + 2\|\beta_{T_0^c}\|_{\ell_1}) \quad (4.11)$$

$$\leq \frac{2\sqrt{2}}{1-\delta}S^{1/2}\lambda_p + \frac{\theta}{1-\delta}\|h_{T_0}\|_{\ell_2} + \frac{2\theta}{1-\delta}S^{-1/2}\|\beta_{T_0^c}\|_{\ell_1} \quad (4.12)$$

Manipulation of this last inequality yields

$$\|h_{T_{01}}\|_{\ell_2} \leq \frac{2\sqrt{2}}{1-\delta-\theta}\lambda_p + \frac{2\theta}{1-\delta-\theta}S^{-1/2}\|\beta_{T_0^c}\|_{\ell_1} \quad (4.13)$$

Combining bounds, we arrive at

$$\|h\|_{\ell_2} \leq \frac{4\sqrt{2}}{1-\delta-\theta}S^{1/2}\lambda_p + 2\frac{1-\delta+\theta}{1-\delta-\theta}S^{-1/2}\|\beta_{T_0^c}\|_{\ell_1} \quad (4.14)$$

$$= \frac{4\sqrt{2}}{1-\delta-\theta}S^{1/2}\lambda_p + 2\sqrt{n}\frac{1-\delta+\theta}{1-\delta-\theta}S^{-1/2}\|\beta_{T_0^c}^0\|_{\ell_1} \quad (4.15)$$

$$\leq \frac{4\sqrt{2}}{1-\delta-\theta}S^{1/2}\lambda_p + 2\frac{1-\delta+\theta}{1-\delta-\theta}S^{-1/2}R \quad (4.16)$$

Scaling by \sqrt{n} then yields the result. \square

To simplify what follows, we set ε equal to the bound in Theorem 2.1 and then define

$$\mathcal{P}_{S,R}^\varepsilon = \{\beta \in \mathbb{R}^p : \|\beta - \beta^0\|_{\ell_2} > 2\varepsilon\} \cap \mathcal{P}_{S,R}, \quad (4.17)$$

which we shall denote as \mathcal{P} when there is no possibility for ambiguity. We are now ready to define the set of y 's which produce controllable denominators, and we also prove the properties we shall exploit.

Lemma 4.1. *Define the critical region $\mathcal{C} = \{y \in \mathbb{R}^n : \|\hat{\beta} - \beta^0\|_{\ell_2} > \varepsilon\}$ and our hypothesis test is then $\Phi(y) = \mathbf{1}_{\mathcal{C}}(y)$. Then,*

1. $\mathbb{E}_{\beta^0} \Phi \leq \frac{1}{p^\alpha \sqrt{\pi \log p}}$
2. $\sup_{\beta \in \mathcal{P}} \mathbb{E}_\beta (1 - \Phi) \leq \frac{1}{p^\alpha \sqrt{\pi \log p}}$

Proof. First, we compute the type I error rate for Φ :

$$\mathbb{E}_{\beta^0} \Phi = \text{pr}_{\beta^0}(\mathcal{C}) \leq \frac{1}{p^\alpha \sqrt{\pi \log p}}. \quad (4.18)$$

In a similar fashion, we have

$$\sup_{\beta \in \mathcal{P}} \mathbb{E}_\beta (1 - \Phi) = \sup_{\beta \in \mathcal{P}} \text{pr}_\beta \{\|\hat{\beta} - \beta^0\|_{\ell_2} \leq \varepsilon\} \quad (4.19)$$

The reverse triangle inequality then yields the bound

$$\sup_{\beta \in \mathcal{P}} \mathbb{E}_\beta (1 - \Phi) \leq \sup_{\beta \in \mathcal{P}} \text{pr}_\beta \{\|\hat{\beta} - \beta\|_{\ell_2} \geq -\varepsilon + \|\beta - \beta^0\|_{\ell_2}\} \quad (4.20)$$

$$\leq \sup_{\beta \in \mathcal{P}} \text{pr}_\beta \{\|\hat{\beta} - \beta\|_{\ell_2} > \varepsilon\} \quad (4.21)$$

where we have used the fact that $\|\beta - \beta^0\|_{\ell_2} > 2\varepsilon$ for all $\beta \in \mathcal{P}$. Moreover, since β is S -sparse for all $\beta \in \mathcal{P}$, $\text{pr}_\beta \{\|\hat{\beta} - \beta\|_{\ell_2} > \varepsilon\} \leq \frac{1}{p^\alpha \sqrt{\pi \log p}}$, and we obtain the desired bound on the supremum. This completes the proof. \square

Because of the behavior of the Dantzig selector, this hypothesis test is only useful for distinguishing between sparse vectors. That is, a large set of non-sparse vectors may trigger a type II error. While this may be a damning indictment for its utility as a practical hypothesis test, we merely employ Φ in the theoretical argument for our main theorem.

Proof of Theorem 3.1. We apply the standard divide-and-conquer strategy originally devised by Schwartz [11] to obtain

$$\Pi(\mathcal{B}|y) = \Phi(y)\Pi(\mathcal{B}|y) + (1 - \Phi(y))\Pi(\mathcal{B}|y)\mathbf{1}_{\mathcal{A}_\kappa}(y) \quad (4.22)$$

$$+ (1 - \Phi(y))\Pi(\mathcal{B}|y)\mathbf{1}_{\mathcal{A}_\kappa^c}(y) \quad (4.23)$$

$$\leq \Phi(y) + (1 - \Phi(y))\Pi(\mathcal{B}|y)\mathbf{1}_{\mathcal{A}_\kappa}(y) + \mathbf{1}_{\mathcal{A}_\kappa^c}(y) \quad (4.24)$$

By Lemma 1, we have that $\mathbb{E}_{\beta^0} \Phi < \frac{1}{p^\alpha \sqrt{\pi \log p}}$, so this term is immediately eliminated. Additionally, we have that $\mathbb{E}_{\beta^0} \mathbf{1}_{\mathcal{A}_\kappa^c}(y) = \text{pr}_{\beta^0}(\mathcal{A}_\kappa^c)$. Having dispensed with the first and third terms, we proceed to attack the middle term.

We first multiply this remaining term by a form of 1 to obtain

$$\frac{(1 - \Phi(y)) \int_{\mathcal{B}} \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta)}{\int \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta)} \mathbf{1}_{\mathcal{A}_\kappa} \quad (4.25)$$

Now, we bound the denominator by the expression

$$\int \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta) \geq \exp\{-\nu \log p\} \Pi(\mathcal{D}_\nu(y)) \quad (4.26)$$

where

$$\mathcal{D}_\nu(y) = \left\{ \beta \in \mathbb{R}^p : \frac{1}{\log p} \log \frac{f(y|\beta^0)}{f(y|\beta)} < \nu \right\} \quad (4.27)$$

$$= \left\{ \beta \in \mathbb{R}^p : \frac{1}{\log p} (\|y - X\beta\|_{\ell_2}^2 - \|y - X\beta^0\|_{\ell_2}^2) < 2\sigma^2\nu \right\} \quad (4.28)$$

$$= \left\{ \beta \in \mathbb{R}^p : \|y - X\beta\|_{\ell_2}^2 - \|y - X\beta^0\|_{\ell_2}^2 < 2\sigma^2\nu \log p \right\} \quad (4.29)$$

By applying the Hölder inequality and the definition of the operator norm, it is easy to see that the left-hand side of the inequality in (4.29) is

$$= \langle (y - X\beta) + (y - X\beta^0), (y - X\beta) - (y - X\beta^0) \rangle \quad (4.30)$$

$$= \langle 2y - 2X\beta^0, X(\beta^0 - \beta) \rangle + \langle X(\beta^0 - \beta), X(\beta^0 - \beta) \rangle \quad (4.31)$$

$$\leq 2\|X^T(y - X\beta^0)\|_{\ell_v} \|\beta - \beta^0\|_{\ell_u} + \|X^T X\|_{\ell_u \rightarrow \ell_v} \|\beta - \beta^0\|_{\ell_u}^2 \quad (4.32)$$

$$\leq 2\kappa \|\beta - \beta^0\|_{\ell_u} + \|X^T X\|_{\ell_u \rightarrow \ell_v} \|\beta - \beta^0\|_{\ell_u}^2 \quad (4.33)$$

since $\kappa > \|X^T(y - X\beta^0)\|_{\ell_v}$ for $y \in \mathcal{A}_\kappa$. Following (4.29), we force $\|\beta - \beta^0\|_{\ell_2}$ to satisfy the inequality

$$\|X^T X\|_{\ell_u \rightarrow \ell_v} \|\beta - \beta^0\|_{\ell_u}^2 + 2\kappa \|\beta - \beta^0\|_{\ell_u} < 2\sigma^2\nu \log p, \quad (4.34)$$

which then yields the bound on $\|\beta - \beta^0\|_{\ell_u}$

$$< \frac{\sqrt{4\kappa^2 + 8\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2\nu \log p} - 2\kappa}{2\|X^T X\|_{\ell_u \rightarrow \ell_v}} \quad (4.35)$$

$$= \left(\frac{\sqrt{8\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2\nu \log p}}{2\kappa + \sqrt{4\kappa^2 + 8\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2\nu \log p}} \right) \frac{\sqrt{8\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2\nu \log p}}{2\|X^T X\|_{\ell_u \rightarrow \ell_v}} \quad (4.36)$$

$$= \left(\frac{\sqrt{2\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2\nu \log p}}{\kappa + \sqrt{\kappa^2 + 2\|X^T X\|_{\ell_u \rightarrow \ell_v} \sigma^2\nu \log p}} \right) \sqrt{\frac{2\sigma^2\nu \log p}{\|X^T X\|_{\ell_u \rightarrow \ell_v}}}. \quad (4.37)$$

Based on this sequence of inequalities, we conclude that $\mathcal{D}_{\nu,\kappa} \subset \mathcal{D}_\nu(y)$ when $y \in \mathcal{A}_\kappa$. Putting this all together, we have that $\Pi(\mathcal{D}_\nu(y)) \geq \Pi(\mathcal{D}_{\nu,\kappa})$ for $y \in \mathcal{A}_\kappa$, and hence

$$\int \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta) \geq p^{-\nu} \Pi(\mathcal{D}_{\nu,\kappa}) \quad (4.38)$$

for all $y \in \mathcal{A}_\kappa$. Applying this, we obtain the bound

$$(1 - \Phi(y))\Pi(\mathcal{B})\mathbf{1}_{\mathcal{A}_\kappa}(y) \leq \frac{(1 - \Phi(y)) \int_{\mathcal{B}} \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta)}{p^{-\nu}\Pi(\mathcal{D}_{\nu,\kappa})}. \quad (4.39)$$

Taking the expectation of the numerator and applying Tonelli yields

$$\mathbb{E}_{\beta^0}(1 - \Phi(y)) \int_{\mathcal{B}} \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta) = \int_{\mathcal{B}} \mathbb{E}_{\beta^0}(1 - \Phi(y)) \frac{f(y|\beta)}{f(y|\beta^0)} d\Pi(\beta) \quad (4.40)$$

$$= \int_{\mathcal{B}} \mathbb{E}_{\beta}(1 - \Phi(y)) d\Pi(\beta) \quad (4.41)$$

We now split this and bound using Lemma 1:

$$\int_{\mathcal{B}} \mathbb{E}_{\beta}(1 - \Phi(y)) d\Pi(\beta) = \int_{\mathcal{B} \setminus \mathcal{P}} \mathbb{E}_{\beta}(1 - \Phi(y)) d\Pi(\beta) \quad (4.42)$$

$$+ \int_{\mathcal{P}} \mathbb{E}_{\beta}(1 - \Phi(y)) d\Pi(\beta) \quad (4.43)$$

$$\leq \Pi(\mathcal{B} \setminus \mathcal{P}) + \Pi(\mathcal{P}) \frac{1}{p^\alpha \sqrt{\pi \log p}} \quad (4.44)$$

$$\leq \Pi(\mathcal{B} \setminus \mathcal{P}) + \frac{1}{p^\alpha \sqrt{\pi \log p}} \quad (4.45)$$

This establishes the result. \square

Appendix B

In order to prove Theorem 4.1, we shall require some additional notation. For a fixed σ, V, X, y, e , and $\gamma \{0, 1\}^p$, we let X_γ denote the matrix obtained by deleting the columns of X with indices i such that $\gamma_i = 0$, P_γ denote the orthogonal projection onto the span of the columns of X_γ ,

$$\Sigma_\gamma = \left(\frac{1}{\sigma^2} X_\gamma^T X_\gamma + \frac{1}{V^2} I_{S \times S} \right)^{-1/2}, \quad (4.46)$$

and

$$\mu_\gamma = \frac{1}{\sigma^2} \Sigma_\gamma^2 X_\gamma^T y. \quad (4.47)$$

Additionally, we shall slightly abuse notation by letting β_γ denote the projection of β onto the coordinates indicated by γ and the Hadamard product of β with γ depending upon the context. Finally, for $\gamma, \gamma' \in \{0, 1\}^p$, we shall write $\gamma \leq \gamma'$ to indicate that γ' dominates γ entry wise.

We first begin with a simple probabilistic noise bound in the spirit of Candès and Tao [5]. The proof is a simple application of the Markov inequality.

Lemma 4.2. *Assuming that $e_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, \dots, n$ and $\|\tilde{X}_i\|_{\ell_2}^2 = 1$ for $i = 1, \dots, p$. If*

$$\mathcal{E} = \{e \in \mathbb{R}^n : \|\tilde{X}_\gamma^T e\|_{\ell_2}^2 \leq 4\sigma^2(1 + \alpha)|\gamma| \log p, \forall \gamma \in \{0, 1\}^p\}, \quad (4.48)$$

then

$$\text{pr}_e(\mathcal{E}) > 1 - \frac{1}{p^\alpha \sqrt{\pi \log p}}. \quad (4.49)$$

The next lemma we shall require deterministically bounds the difference between similar operators.

Lemma 4.3. *Assume (A1)-(A4), if $\gamma \in \{0, 1\}^p$ satisfies $\gamma^0 \leq \gamma$ and $|\gamma| \leq 2S$, we have that*

$$\left\| P_\gamma - X_\gamma(X_\gamma^T X_\gamma + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_\gamma^T \right\|_{\ell_2 \rightarrow \ell_2} \leq \frac{\sigma^2}{n(1 - \delta)V^2 + \sigma^2}. \quad (4.50)$$

and

$$\left\| I_{|\gamma| \times |\gamma|} - (X_\gamma^T X_\gamma + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_\gamma^T X_\gamma \right\|_{\ell_2 \rightarrow \ell_2} \leq \frac{\sigma^2}{n(1 - \delta)V^2 + \sigma^2}. \quad (4.51)$$

We shall also require bounds on the determinants of the restricted operators. This lemma and the preceding lemma follow from the RIP hypothesis and application of an SVD.

Lemma 4.4. *Assuming (A1) and (A4), if $\gamma \in \{0, 1\}^p$ satisfies $|\gamma| \leq 2S$, then we have that*

$$\left(\frac{n(1 + \delta)}{\sigma^2} + \frac{1}{V^2} \right)^{-|\gamma|/2} \leq \det(\Sigma_\gamma) \leq \left(\frac{n(1 - \delta)}{\sigma^2} + \frac{1}{V^2} \right)^{-|\gamma|/2}. \quad (4.52)$$

On the other hand, if $|\gamma| > 2S$, then

$$\det(\Sigma_\gamma) \leq \left(\frac{n(1 - \delta)}{\sigma^2} + \frac{1}{V^2} \right)^{-S}. \quad (4.53)$$

Now, we exhibit a bound on the difference between norms of different reconstructions.

Lemma 4.5. *Assume (A1), (A2), and (A4). If $\gamma, \gamma' \in \{0, 1\}^p$ satisfy $\gamma^0 \leq \gamma$ and $|\gamma'| \leq 2S$, then*

$$(X\beta^0)^T (P_{\gamma'} - P_\gamma) X\beta^0 \leq -n \left(1 - \delta - \frac{\theta^2}{1 - \delta} \right) \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2. \quad (4.54)$$

Moreover, $\delta + \frac{\theta^2}{1 - \delta} < 1$.

Proof. Since $\gamma^0 \leq \gamma$, we have that $P_\gamma X \beta^0 = X_{\gamma^0} \beta_{\gamma^0}^0$, and therefore

$$\begin{aligned} (X \beta^0)^T (P_{\gamma'} - P_\gamma) X \beta^0 &= (X_{\gamma^0} \beta_{\gamma^0}^0)^T (P_{\gamma'} - I_{n \times n}) X_{\gamma^0} \beta_{\gamma^0}^0 \\ &= \left(X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0 \right)^T (P_{\gamma'} - I_{n \times n}) X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0 \\ &= -\|X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 + \|P_{\gamma'} X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2. \end{aligned} \quad (4.55)$$

We then have

$$\|P_{\gamma'} X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 \leq \frac{n}{1-\delta} \|\tilde{X}_{\gamma'}^T \tilde{X}_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 \quad (4.56)$$

$$\leq n \frac{\theta^2}{1-\delta} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 \quad (4.57)$$

since

$$\|\tilde{X}_{\gamma'}^T \tilde{X}_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 = \langle \tilde{X}_{\gamma'} \tilde{X}_{\gamma'}^T \tilde{X}_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0, \tilde{X}_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0 \rangle \quad (4.58)$$

$$\leq \theta \|\tilde{X}_{\gamma'}^T \tilde{X}_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2} \quad (4.59)$$

implies $\|\tilde{X}_{\gamma'}^T \tilde{X}_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 \leq \theta^2 \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2$. Combining this with the fact that $n(1-\delta) \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 \leq \|X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2$, we obtain the bound

$$(X \beta^0)^T (P_{\gamma'} - P_\gamma) X \beta^0 \leq -n \left(1 - \delta - \frac{\theta^2}{1-\delta} \right) \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2. \quad (4.60)$$

Finally, note that

$$\theta + \delta < 1 \implies \theta^2 < (1-\delta)^2 \implies \frac{\theta^2}{1-\delta} \leq 1-\delta \implies \delta + \frac{\theta^2}{1-\delta} < 1. \quad (4.61)$$

□

This next lemma bounds the differences of inner products of reconstructions with the noise vector.

Lemma 4.6. *Assume (A1) through (A4), and $e \in \mathcal{E}$ from (4.48). If $\gamma^0 \leq \gamma$ and $|\gamma'| \leq 2S$, then*

$$|(X \beta^0)^T (P_{\gamma'} - P_\gamma) e| \leq 2\sigma \frac{1-\delta+\theta}{1-\delta} \sqrt{(1+\alpha) \max\{|\gamma^0 \setminus \gamma'|, |\gamma'|\} n \log p} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}.$$

Proof. We compute

$$|(X \beta^0)^T (P_{\gamma'} - P_\gamma) e| = |(X_{\gamma^0} \beta_{\gamma^0}^0)^T (P_{\gamma'} - I_{n \times n}) e| \quad (4.62)$$

$$= |(X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0)^T (P_{\gamma'} - I_{n \times n}) e| \quad (4.63)$$

$$= |(\beta_{\gamma^0 \setminus \gamma'}^0)^T X_{\gamma^0 \setminus \gamma'}^T e + (X_{\gamma^0 \setminus \gamma'} \beta_{\gamma^0 \setminus \gamma'}^0)^T P_{\gamma'} e| \quad (4.64)$$

$$\leq \|X_{\gamma^0 \setminus \gamma'}^T e\|_{\ell_2} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2} + \|P_{\gamma'} X_{\gamma^0 \setminus \gamma'}^T e\|_{\ell_2} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}$$

$$\leq \|X_{\gamma^0 \setminus \gamma'}^T e\|_{\ell_2} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2} + \frac{\theta}{1-\delta} \|X_{\gamma'}^T e\|_{\ell_2} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}$$

$$\leq 2\sigma \left(1 + \frac{\theta}{1-\delta} \right) \sqrt{(1+\alpha) n \max\{|\gamma^0 \setminus \gamma'|, |\gamma'|\} \log p} \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}.$$

□

Our last lemma is used to clean up a calculation that arises.

Lemma 4.7. *Let*

$$\tau = A \sqrt{\frac{(1+\alpha)S \log p}{n}} = \frac{8\sqrt{2}\sigma}{1-\delta-\theta} \sqrt{\frac{(1+\alpha)S \log p}{n}}. \quad (4.65)$$

Then

$$-n \left(1 - \delta + \frac{\theta^2}{1-\delta} \right) \tau^2 + 4\sqrt{2}\sigma \frac{1-\delta+\theta}{1-\delta} \sqrt{(1+\alpha)Sn \log p} \tau \leq -\frac{n}{2} \left(1 - \delta + \frac{\theta^2}{1-\delta} \right) \tau^2.$$

Proof of Theorem 4.1. First, we exhibit an explicit formula for the posterior. Given any measurable set $U \subset \mathbb{R}^n$, we have that

$$\begin{aligned} & \int_U f(y|\beta) d\Pi(\beta) \quad (4.66) \\ &= \binom{p}{S}^{-1} \sum_{\gamma \in \{0,1\}_S^p} \int_U (2\pi\sigma^2)^{-n/2} e^{-\|y-X\beta_\gamma\|_{\ell_2}^2/2\sigma^2} (2\pi V^2)^{-S/2} e^{-\|\beta_\gamma\|_{\ell_2}^2/2\tau^2} d\beta_\gamma \\ &= \binom{p}{S}^{-1} (2\pi\sigma^2)^{-n/2} (2\pi V^2)^{-S/2} \sum_{\gamma \in \{0,1\}_S^p} \int_U e^{-\frac{\|y-X\beta_\gamma\|_{\ell_2}^2}{2\sigma^2} - \frac{\|\beta_\gamma\|_{\ell_2}^2}{2V^2}} d\beta_\gamma \quad (4.67) \end{aligned}$$

Completing the square gives us

$$\frac{\|y-X\beta_\gamma\|_{\ell_2}^2}{2\sigma^2} + \frac{\|\beta_\gamma\|_{\ell_2}^2}{2V^2} = \frac{1}{2}(\beta_\gamma - \mu_\gamma)^T \Sigma_\gamma^{-2} (\beta_\gamma - \mu_\gamma) + \frac{1}{2\sigma^2} \|y\|_{\ell_2}^2 - \frac{1}{2} \mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma$$

Therefore, if $U = \mathbb{R}^n$, we have that

$$\int_U e^{-\frac{\|y-X\beta_\gamma\|_{\ell_2}^2}{2\sigma^2} - \frac{\|\beta_\gamma\|_{\ell_2}^2}{2V^2}} d\beta_\gamma = (2\pi)^{S/2} \det(\Sigma_\gamma) e^{-\frac{1}{2\sigma^2} \|y\|_{\ell_2}^2 + \frac{1}{2} \mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma} \quad (4.68)$$

and hence

$$\int_{\mathbb{R}^n} f(y|\beta) d\Pi(\beta) = e^{-\frac{1}{2\sigma^2} \|y\|_{\ell_2}^2} \binom{p}{S}^{-1} (2\pi\sigma^2)^{-n/2} V^{-S} \sum_{\gamma \in \{0,1\}_S^p} \det(\Sigma_\gamma) e^{\frac{1}{2} \mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}$$

On the other hand,

$$\begin{aligned} & \int_{B_{2\varepsilon}^{\ell_2}(\beta^0)} f(y|\beta) d\Pi(\beta) \quad (4.69) \\ &= \binom{p}{S}^{-1} (2\pi\sigma^2)^{-n/2} (2\pi V^2)^{-S/2} \sum_{\gamma \in \{0,1\}_S^p} \int_{B_{2\varepsilon}^{\ell_2}(\beta^0)} e^{-\frac{\|y-X\beta_\gamma\|_{\ell_2}^2}{2\sigma^2} - \frac{\|\beta_\gamma\|_{\ell_2}^2}{2V^2}} d\beta_\gamma \\ &= e^{-\frac{1}{2\sigma^2} \|y\|_{\ell_2}^2} \binom{p}{S}^{-1} (2\pi\sigma^2)^{-n/2} V^{-S} \sum_{\gamma \in \{0,1\}_S^p} \det(\Sigma_\gamma) e^{\frac{1}{2} \mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma} \int_{B_{2\varepsilon}^{\ell_2}(\beta^0)} \frac{e^{-\frac{1}{2}(\beta_\gamma - \mu_\gamma)^T \Sigma_\gamma^{-2} (\beta_\gamma - \mu_\gamma)}}{\sqrt{2\pi}^S \det(\Sigma_\gamma)} d\beta_\gamma \end{aligned}$$

Putting this all together, we have

$$\begin{aligned}
 \Pi(B_{2\varepsilon}^{\ell_2}(\beta^0)|y) &= \frac{\int_{B_{2\varepsilon}^{\ell_2}(\beta^0)} f(y|\beta) d\Pi(\beta)}{\int_{\mathbb{R}^n} f(y|\beta) d\Pi(\beta)} \\
 &= \frac{\sum_{\gamma \in \{0,1\}_S^p} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma} \int_{B_{2\varepsilon}^{\ell_2}(\beta^0)} \frac{e^{-\frac{1}{2}(\beta_\gamma - \mu_\gamma)^T \Sigma_\gamma^{-2} (\beta_\gamma - \mu_\gamma)}}{\sqrt{2\pi}^S \det(\Sigma_\gamma)} d\beta_\gamma}{\sum_{\gamma \in \{0,1\}_S^p} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}}.
 \end{aligned} \tag{4.70}$$

Now that we have an explicit expression for the posterior, we bound this expression below by reducing the sum in the numerator to the indices in the set

$$G = \left\{ \gamma \in \{0,1\}_S^p : \|\beta_{\gamma^0 \setminus \gamma}\|_{\ell_2} \leq A \sqrt{\frac{(1+\alpha)S \log p}{n}} \right\} \tag{4.71}$$

That is, we restrict to the indices that capture most of the mass of β^0 . If $\gamma \in G$, then

$$\|\beta_\gamma^0 - \mu_\gamma\|_{\ell_2} = \|\beta_\gamma^0 - (X_\gamma^T X_\gamma + \frac{\sigma^2}{V^2} I_{n \times n})^{-1} X_\gamma^T (X \beta^0 + e)\|_{\ell_2} \tag{4.72}$$

$$\leq \|(I_{S \times S} - (X_\gamma^T X_\gamma + \frac{\sigma^2}{V^2} I_{n \times n})^{-1} X_\gamma^T X_\gamma) \beta_\gamma^0\|_{\ell_2} \tag{4.73}$$

$$+ \|(X_\gamma^T X_\gamma + \frac{\sigma^2}{V^2} I_{n \times n})^{-1} X_\gamma^T X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0\|_{\ell_2} \tag{4.74}$$

$$+ \|(X_\gamma^T X_\gamma + \frac{\sigma^2}{V^2} I_{n \times n})^{-1} X_\gamma^T e\|_{\ell_2} \tag{4.75}$$

$$\leq \frac{\sigma^2}{n(1-\delta)V^2 + \sigma^2} \|\beta_\gamma^0\|_{\ell_2} \tag{4.76}$$

$$+ \frac{n\theta}{n(1-\delta) + \sigma^2/V^2} \|\beta_{\gamma^0 \setminus \gamma}^0\|_{\ell_2} \tag{4.77}$$

$$+ 2 \frac{\sqrt{nS(1+\alpha)\sigma^2 \log p}}{n(1-\delta) + \sigma^2/V^2} \tag{4.78}$$

$$= \frac{\sigma^2 \|\beta_\gamma^0\|_{\ell_2} + n\theta V^2 \|\beta_{\gamma^0 \setminus \gamma}^0\|_{\ell_2} + 2\sigma V^2 \sqrt{(1+\alpha)Sn \log p}}{n(1-\delta)V^2 + \sigma^2}.$$

Given this bound, we may conclude that (using the bound $\|\beta_\gamma^0\|_{\ell_2} \leq C\sqrt{S}$, but note that a large enough V may be chosen to dampen this contribution)

$$\|\beta_\gamma^0 - \mu_\gamma\|_{\ell_2} \leq \|\beta_{\gamma^0 \setminus \gamma}^0\|_{\ell_2} + \|\beta_\gamma^0 - \mu_\gamma\|_{\ell_2} \tag{4.79}$$

$$\leq \frac{\sigma^2 \|\beta_\gamma^0\|_{\ell_2} + (n(1-\delta) + \theta)V^2 + \sigma^2 \|\beta_{\gamma^0 \setminus \gamma}^0\|_{\ell_2} + 2\sigma V^2 \sqrt{(1+\alpha)Sn \log p}}{n(1-\delta)V^2 + \sigma^2}$$

$$\leq \varepsilon \tag{4.80}$$

and hence

$$\begin{aligned}
 \int_{B_{2\varepsilon}^{\ell_2}(\beta^0)} \frac{e^{-\frac{1}{2}(\beta_\gamma - \mu_\gamma)^T \Sigma_\gamma^{-2}(\beta_\gamma - \mu_\gamma)}}{\sqrt{2\pi}^S \det(\Sigma_\gamma)} d\beta_\gamma &\geq \int_{B_\varepsilon^{\ell_2}(\mu_\gamma)} \frac{e^{-\frac{1}{2}(\beta_\gamma - \mu_\gamma)^T \Sigma_\gamma^{-2}(\beta_\gamma - \mu_\gamma)}}{\sqrt{2\pi}^S \det(\Sigma_\gamma)} d\beta_\gamma \\
 &\geq \int_{B_\varepsilon^{\ell_2}(\mu_\gamma)} \frac{e^{-\frac{1}{2}(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2})\|\beta_\gamma - \mu_\gamma\|_{\ell_2}^2}}{\sqrt{2\pi} \left(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2}\right)^{-1} S} d\beta_\gamma \\
 &\geq 1 - e^{-\frac{1}{4}(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2})\varepsilon^2}. \tag{4.81}
 \end{aligned}$$

This last expression holds because of the hypothesis $(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2})\varepsilon^2 \geq S/2$. At this stage, we have constructed the bound

$$\Pi(B_{2\varepsilon}^{\ell_2}(\beta^0)|y) \geq \frac{\sum_{\gamma \in G} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}}{\sum_{\gamma \in \{0,1\}_S^p} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}} (1 - e^{-\frac{1}{4}(n\frac{1-\delta}{\sigma^2} + \frac{1}{V^2})\varepsilon^2}). \tag{4.82}$$

We now approach the expression

$$\begin{aligned}
 &\frac{\sum_{\gamma \in G} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}}{\sum_{\gamma \in \{0,1\}_S^p} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}} \tag{4.83} \\
 &= \frac{1}{1 + \frac{\sum_{\gamma \in \{0,1\}_S^p \setminus G} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}}{\sum_{\gamma \in G} \det(\Sigma_\gamma) e^{\frac{1}{2}\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma}}} \\
 &= \frac{1}{1 + \sum_{\gamma' \in \{0,1\}_S^p \setminus G} \frac{1}{\frac{\det(\Sigma_\gamma)}{\det(\Sigma_{\gamma'})} e^{\frac{1}{2}(\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma - \mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'})}}} \\
 &\geq \frac{1}{1 + \sum_{\gamma' \in \{0,1\}_S^p \setminus G} \frac{1}{\sum_{\gamma^0 \leq \gamma} \frac{\det(\Sigma_\gamma)}{\det(\Sigma_{\gamma'})} e^{\frac{1}{2}(\mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma - \mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'})}}}.
 \end{aligned}$$

In the last step, we have reduced the index set over the sum inside the continued fraction from G to its subset $\{\gamma \in \{0,1\}_S^p : \gamma^0 \leq \gamma\}$. Based on this initial bound, we shall seek upper bounds on the expressions

$$\frac{\det(\Sigma_{\gamma'})}{\det(\Sigma_\gamma)} e^{\frac{1}{2}(\mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'} - \mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma)} \tag{4.84}$$

for all $\gamma^0 \leq \gamma$ and $\gamma' \in \{0,1\}_S^p \setminus G$. First, we note that

$$\frac{\det(\Sigma_{\gamma'})}{\det(\Sigma_\gamma)} \leq \left(\frac{n(1+\delta)V^2 + \sigma^2}{n(1-\delta)V^2 + \sigma^2} \right)^{S/2} \tag{4.85}$$

by Lemma 4.4. The remaining expressions that we must examine have the form

$$\exp \left\{ \frac{1}{2} \left(\mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'} - \mu_\gamma^T \Sigma_\gamma^{-2} \mu_\gamma \right) \right\}. \tag{4.86}$$

Since

$$\mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'} - \mu_{\gamma}^T \Sigma_{\gamma}^{-2} \mu_{\gamma} = \left(\frac{1}{\sigma^2} \Sigma_{\gamma'}^2 X_{\gamma'}^T y \right)^T \Sigma_{\gamma'}^{-2} \left(\frac{1}{\sigma^2} \Sigma_{\gamma'}^2 X_{\gamma'}^T y \right) \quad (4.87)$$

$$\begin{aligned} & - \left(\frac{1}{\sigma^2} \Sigma_{\gamma}^2 X_{\gamma}^T y \right)^T \Sigma_{\gamma}^{-2} \left(\frac{1}{\sigma^2} \Sigma_{\gamma}^2 X_{\gamma}^T y \right) \quad (4.88) \\ & = \frac{1}{\sigma^2} y^T X_{\gamma'} (X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{V^2} I_{|\gamma'| \times |\gamma'|})^{-1} X_{\gamma'}^T y \\ & \quad - \frac{1}{\sigma^2} y^T X_{\gamma} (X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_{\gamma}^T y, \end{aligned}$$

we focus on bounding the expression

$$y^T \left(X_{\gamma'} (X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{V^2} I_{|\gamma'| \times |\gamma'|})^{-1} X_{\gamma'}^T - X_{\gamma} (X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_{\gamma}^T \right) y$$

For all $\gamma \in \{0, 1\}^p$, let P_{γ} be the orthogonal projection onto $\text{span}\{X_i : \gamma_i = 1\}$. Then

$$\begin{aligned} & X_{\gamma'} (X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{V^2} I_{|\gamma'| \times |\gamma'|})^{-1} X_{\gamma'}^T - X_{\gamma} (X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_{\gamma}^T \\ & \preceq P_{\gamma'} - P_{\gamma} + \left(P_{\gamma} - X_{\gamma} (X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_{\gamma}^T \right), \end{aligned} \quad (4.89)$$

in the positive definite ordering. By Lemma 4.3 we have

$$y^T \left(P_{\gamma} - X_{\gamma} (X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{V^2} I_{|\gamma| \times |\gamma|})^{-1} X_{\gamma}^T \right) y \quad (4.90)$$

$$\leq \frac{\sigma^2}{n(1-\delta)V^2 + \sigma^2} \|y\|_{\ell_2}^2 \quad (4.91)$$

On the other hand, we may expand

$$y^T (P_{\gamma'} - P_{\gamma}) y = (X_{\gamma_0} \beta_{\gamma_0}^0)^T (P_{\gamma'} - P_{\gamma}) X_{\gamma_0} \beta_{\gamma_0}^0 \quad (4.92)$$

$$+ 2 (X_{\gamma_0} \beta_{\gamma_0}^0)^T (P_{\gamma'} - P_{\gamma}) e \quad (4.93)$$

$$+ e^T (P_{\gamma'} - P_{\gamma}) e, . \quad (4.94)$$

By applying Lemmas 4.5, 4.6, and 4.2 to (4.92), (4.93), and (4.94) respectively, we obtain the bound

$$\begin{aligned} y^T (P_{\gamma'} - P_{\gamma}) y & \leq -n \left(1 - \delta - \frac{\theta^2}{1 - \delta} \right) \|\beta_{\gamma_0 \setminus \gamma'}^0\|_{\ell_2}^2 \\ & \quad + 4\sigma \left(1 + \frac{\theta}{1 - \delta} \right) \sqrt{(1 + \alpha) S n \log p} \|\beta_{\gamma_0 \setminus \gamma'}^0\|_{\ell_2} \\ & \quad + 4 \frac{\sigma^2}{1 - \delta} (1 + \alpha) S \log p \end{aligned}$$

Since $\gamma' \in \{0, 1\}_S^p \setminus G$, we may now employ Lemma 4.7 and the fact that $\gamma' \in \{0, 1\}_S^p \setminus G$ to obtain the bounds

$$\begin{aligned} y^T(P_{\gamma'} - P_{\gamma})y &\leq -\frac{n}{2} \left(1 - \delta - \frac{\theta^2}{1 - \delta}\right) \|\beta_{\gamma^0 \setminus \gamma'}^0\|_{\ell_2}^2 + 4\frac{\sigma^2}{1 - \delta}(1 + \alpha)S \log p \\ &\leq \underbrace{\left(-\frac{A^2}{2} \left(1 - \delta - \frac{\theta^2}{1 - \delta}\right) + 4\frac{\sigma^2}{1 - \delta}\right)}_{=-2\sigma^2(1+\eta)} (1 + \alpha) S \log p. \end{aligned}$$

Accumulating the bounds we have exhibited thus far, we have

$$\frac{1}{2} \left(\mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'} - \mu_{\gamma}^T \Sigma_{\gamma}^{-2} \mu_{\gamma} \right) \leq -(1 + \eta)S \log p + \frac{1}{2} \frac{1}{n(1 - \delta)V^2 + \sigma^2} \|y\|_{\ell_2}^2,$$

and hence

$$\frac{\det(\Sigma_{\gamma'})}{\det(\Sigma_{\gamma})} e^{\frac{1}{2}(\mu_{\gamma'}^T \Sigma_{\gamma'}^{-2} \mu_{\gamma'} - \mu_{\gamma}^T \Sigma_{\gamma}^{-2} \mu_{\gamma})} \leq \left(\frac{n(1 + \delta)V^2 + \sigma^2}{n(1 - \delta)V^2 + \sigma^2} \right)^{S/2} e^{\frac{1}{n(1 - \delta)V^2 + \sigma^2} \frac{\|y\|_{\ell_2}^2}{2}} p^{-(1 + \eta)S}$$

for all $\gamma^0 \leq \gamma$ and $\gamma' \in \{0, 1\}_S^p$. Therefore, we may bound (4.83) from below by

$$\begin{aligned} &\frac{1}{1 + \left(\frac{n(1 + \delta)V^2 + \sigma^2}{n(1 - \delta)V^2 + \sigma^2} \right)^{S/2} e^{\frac{1}{n(1 - \delta)V^2 + \sigma^2} \frac{\|y\|_{\ell_2}^2}{2}} p^{-(1 + \eta)S} |\{0, 1\}_S^p \setminus G|} \\ &\geq \frac{1}{1 + \left(\frac{n(1 + \delta)V^2 + \sigma^2}{n(1 - \delta)V^2 + \sigma^2} \right)^{S/2} e^{\frac{1}{n(1 - \delta)V^2 + \sigma^2} \frac{\|y\|_{\ell_2}^2}{2}} p^{-(1 + \eta)S} \binom{p}{S}} \\ &\geq \frac{1}{1 + \left(e^{\frac{2n(1 + \delta)V^2 + \sigma^2}{n(1 - \delta)V^2 + \sigma^2}} \right)^{S/2} e^{\frac{1}{n(1 - \delta)V^2 + \sigma^2} \frac{\|y\|_{\ell_2}^2}{2}} S^{-S} p^{-\eta S}} \end{aligned}$$

using the fact that $\binom{p}{S} \leq \left(\frac{ep}{S}\right)^S$. This completes the proof. \square

Acknowledgements

This work was partially funded by the Mathematics of Sensing, Exploitation, and Execution (MSEE) program (managed by Dr. Tony Falcone), by the National Science Foundation under grant DMS-1045153, and by grant R01ES17436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). The authors would also like to thank Mauro Maggioni for helpful discussions.

References

- [1] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723.

- [2] BONTEMPS, D. (2011). Bernstein-Von Mises theorems for Gaussian regression with increasing number of regressors. *Annals of Statistics* **39**, 2557–2584.
- [3] CANDÈS, E. J. and PLAN, Y. (2007) Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics* **37**, 2145–2177.
- [4] CANDÈS, E. J. , ROMBERG, J., and TAO, T. (2005). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223.
- [5] CANDÈS, E. J. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Annals of Statistics*, **35** 2313–2351.
- [6] CEVHER, V. (2009). Learning with Compressible Priors. *Proc. Neural Information Processing Systems*, Vancouver, B.C., Canada.
- [7] DONOHO, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, **52**, 1289–1306.
- [8] GHOSAL, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, **5** 315–331.
- [9] GRIBONVAL, R., CEVHER, V. and DAVIES, M. (2011) Compressible priors for high-dimensional statistics, submitted to *IEEE Transactions on Information Theory*.
- [10] JIANG, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *The Annals of Statistics*, **35** 1487–1511.
- [11] SCHWARTZ, L. (1965). On Bayes procedure. *Probability Theory and Related Fields*. **4** 10–26.
- [12] TROPP, J. (2004). *Topics in Sparse Approximation*. Ph.D. dissertation, Univ. Texas at Austin.
- [13] WANG, H.S. (2009). Forward regression for ultra high-dimensional variable screening. *Journal of the American Statistical Association* **104** 1512–1524
- [14] WANG, T. and ZHU, L.X. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, **102** 1141–1151.
- [15] ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *The Annals of Statistics*, **37** 2109–2144.